Chapter 19

AUTOMATIC ACQUISITION OF GL RESOURCES, USING AN EXPLANATORY, SYMBOLIC TECHNIQUE

Vincent Claveau[1] and Pascale Sébillot[2]
[1]OLST, University of Montréal, C.P. 6128, succ. Centre-Ville, Montréal, QC, H3C 3J7, Canada; [2]IRISA, Campus de Beaulieu, 35042 Rennes cedex, France

Abstract: This chapter presents a symbolic machine learning method that automatically infers, from descriptions of noun-verb pairs found in a corpus in which the verb plays (or not) one of the qualia roles of the noun, corpus-specific morpho-syntactic and semantic patterns that convey qualia relations. The patterns are explanatory and linguistically motivated, and can be applied to a corpus to efficiently extract GL resources and populate Generative Lexicons. The linguistic relevance of these patterns is examined, and the N-V qualia pairs that they can detect or not is discussed. Comparisons to other methods for corpus-based qualia couple extraction are also presented.

Key words: automatic GL resource extraction, symbolic machine learning acquisition

1. INTRODUCTION

The Generative Lexicon (GL) theory (Pustejovsky, 1995) has proved its usefulness in the analysis of numerous linguistic phenomena across languages. Moreover, elements from Generative Lexicons have been shown to be relevant in several natural language processing (NLP) applications (*e.g.* information retrieval, *etc.*). For instance, the qualia structure gives access to relational information, crucial for such applications. In particular,

the qualia roles (namely the telic, agentive, constitutive and formal roles) express, in terms of predicative formulae, the basic features of the semantics of nouns. In a GL model, the noun is linked not only to other nouns via traditional lexical relations (such as meronymy and hyperonymy) but also to verbs. For example, the noun *book* is linked to the verbal predicate *read* via its telic role and to the predicate *write* via its agentive role. Hereafter, a noun(N)-verb(V) pair in which V expresses one of the qualia roles of N (like *book-read* or *book-write*) is called a *qualia pair*. Previous work by Fabre and Sébillot (1999) has demonstrated that these N-V relations provide lexical resources that are found to be useful for information retrieval systems. Different studies (Grefenstette, 1997; Pustejovsky *et al.*, 1997, *inter alia*) also show that N-V pairs can feed indexes that help a user to select the most interesting occurrences of a given noun in a text. Moreover, a short survey (Vandenbroucke, 2000) at the documentation center of the Banque Bruxelles Lambert (Brussels) shows that verbs that express a qualia relation seem to be more relevant than others for a document retrieval task; indeed, in this study, no non-qualia N-V pairs were considered as interesting by the documentalists. Furthermore, the global relevance of qualia verbs for the interpretation of binominal sequences (Fabre, 1996) gives access to various interesting applications in the domain of term variations.

Thus, possessing such GL resources is fundamental for many NLP applications. However, there are two main difficulties to handle:

1. the lack of Generative Lexicons or lexical resources containing those qualia pairs;
2. and the fact that verbs in those pairs may vary considerably from one domain to another (especially in technical domains).

A corpus-based method to acquire such N-V qualia resources has to be found, which would eventually lead to an automatic way to populate Generative Lexicons. This is the precise focus of this chapter, in which we propose and describe such a technique.

This chapter is divided in four parts: we first position our acquisition method within the wide domain of corpus-based acquisition techniques for lexical semantic relations, and differentiate it from other attempts to automatically fill in Generative Lexicons. Our approach relies on a symbolic machine-learning method that infers morpho-syntactic and semantic patterns from examples and counter-examples of N-V qualia pairs in context. These patterns characterize the examples from the counter-examples and then can be applied on a corpus in order to retrieve new N-V qualia pairs. The second part of the text is dedicated to the presentation of our symbolic learning tool, named ASARES, and the description of the corpus on which it has been trained and evaluated. One of the interests for choosing a symbolic method is to obtain explicative patterns, *i.e.* patterns that explain the concept of

qualia role as it is expressed in the studied corpus. The third section presents the inferred patterns, and discusses their linguistic relevance. Finally, a complete evaluation of ASARES is provided in terms of correct N-V qualia pair extraction, and we compare its acquisition performances to those of standard statistical and syntactical approaches. The linguistic discussions in those two last parts of the text are based on a work jointly realized with P. Bouillon (ISSCO, Geneva, Switzerland) and C. Fabre (ERSS, Toulouse, France).

2. AUTOMATIC ACQUISITION OF SEMANTIC RELATIONS

Numerous studies have been dedicated to the corpus-based acquisition of semantic relations. Grefenstette (1994) and Pichon and Sébillot (1997) provide some states-of-the-art of the domain, and Manning and Schütze (1999) describe a large panel of statistical methods that have been used for that purpose. Rather than an exhaustive description of all the elaborated techniques, we present here a reading of the domain, structured by the type of global approach that they can choose. We then give some arguments explaining our choice of a symbolic technique to acquire N-V qualia pairs, and conclude this section by an overview of the (few) studies that have already been realized about Generative Lexicon filling.

2.1 Overview of possible methods

One relevant way to structure the domain of lexical relation acquisition from corpora is to oppose numerical *versus* symbolic approaches. Numerical approaches of acquisition exploit the frequential aspect of data while symbolic approaches exploit the structural aspect of data, and use symbolic information. Note that no assumption is made about the actual technique manipulating symbolic or numerical information; a statistical technique can be used to acquire lexical relations on the basis of symbolic information, and conversely, a symbolic technique can make the most of numerical information.

Within the numerical approach, relations between lexical units can be acquired by studying word cooccurrences in a text window (or specific syntactic structures). The strength of the association is usually evaluated with the help of a statistical score (association coefficient) that detect words appearing together in a statistically significant way. For example, Church and Hanks's work (1989) is based on such a statistical cooccurrence method.

Following the linguistic principles of Harris (1989), numerical distributional analysis methods respect a 3-step approach: extraction of the cooccurrents of one word (within a text window or a syntactic context), evaluation of proximity/distance between two terms, based on their shared or not shared cooccurrents (various measures are defined), clustering into classes (*e.g.* following different data analysis or graph techniques). For example, Bouaud *et al*. (1997) and Grefenstette (1994) follow this kind of technique to discover paradigmatic relations.

The symbolic approach of acquisition groups into two strategies: symbolic linguistic approach, and machine-learning (ML) approach. In the first one, operational definitions of the elements to acquire are manually established by linguists, usually in the form of morpho-lexical patterns that carry the relations that are studied, or by a list of linguistic clues (*e.g.* see Oueslati, 1999). However, when such patterns or clues are unknown, but examples of elements respecting the target relation are known, ML can be used to automatically extract patterns from the descriptions of those examples. The technique is based on a 5-step methodology initiated by Hearst (1992):

1. select one target relation R;
2. gather a list of pairs following relation R;
3. find the sentences that contain those pairs; keep their lexical and syntactic contexts;
4. detect common points between those contexts; suppose that they form a pattern for R;
5. apply the patterns to get new pairs and go back to 3.

ML (inductive logic programming, grammatical inference, *etc.*) (Mitchell, 1997) offers a framework to automate step 4, and aims at automatically producing unknown morpho-lexical patterns that carry the target relation.

Both approaches present advantages and drawbacks. Numerical approaches are usually portable and automatic but produce non-interpretable results; the detection is realized at the corpus level: thus, the detection of one specific occurrence cannot be explained. Symbolic approaches need *a priori* knowledge (patterns or examples), but produce interpretable results; detection is done at the occurrence level.

2.2 Arguments to choose a symbolic ML technique

Let us have a look at each kind of methods listed above and examine its relevance for a corpus-based acquisition of N-V qualia pairs.

First, a N-V qualia pair can be considered as a special kind of cooccurrence, and a statistical approach that extracts N-V pairs related in a

statistically significant way can be chosen. We have however proved that this type of methods is not accurate enough to extract precise relations, *i.e.* N-V pairs linked by a qualia relation *versus* other pairs in our case (see Section 5.3).

Another possibility is to use a symbolic linguistic approach and to extract the N-V pairs by spotting a set of syntactic structures related to qualia roles. In this last case, the advantage is that such patterns can be very precise, but the major problem is that the patterns that carry N-V qualia pairs in a given corpus are mainly unknown; they have to be defined, and adapted to every new text and corpus.

In our case, we have no *a priori* knowledge concerning the structures that are likely to convey qualia roles in a corpus, but we are able to (manually) find some examples of N-V qualia pairs in a text to feed an automatic technique. Thus, we have developed and applied a supervised symbolic machine-learning method. This method automatically produces general rules that explain what, in terms of surrounding context (part-of-speech and semantic tags; see Section 3) in a text, characterizes examples of N-V qualia pairs from non-qualia ones in a given corpus. The rules produced this way—morpho-syntactic and semantic patterns—are then applied to the corpus to exhibit unseen qualia N-V pairs. Therefore, with this system, we aim at combining the precision of linguistic rules (or patterns) in extraction tasks and the flexibility of an automated method. Unlike most statistical methods that only provide a predictor (this N-V pair is qualia, this one is not), our symbolic ML method infers general rules able to explain the examples, that is, bring relevant and linguistically interpretable elements about the predicted qualia relations in the studied corpus.

2.3 Related work

Only a few projects have been undertaken to automatically construct qualia structures. Among them, Pustejovsky *et al.* (1993) propose to acquire elements of these structures from a syntactically-tagged corpus by the means of a cooccurrence-based statistical extraction technique coupled with a set of heuristics, *i.e.* syntactic patterns. However, no precise evaluation of the performances of this work is given and this study makes strong assumptions on the structures conveying the qualia relations and heavily relies on the good results of syntactic parsers, not available for most of the languages.

Using qualia verbs of nouns to define a framework for logical metonymy interpretation, Lapata and Lascarides (2003) also present an acquisition method for N-V qualia pairs. This technique relies on a probabilistic learning based on Naive Bayes (Mitchell, 1997), and uses a syntactic parser to

establish the necessary joint appearance probabilities. More than on the extraction task itself, the evaluation of this work mostly focuses on the possibilities of acquired N-V pairs to interpret metonymies. As the previous one, this study also uses a syntactic parser, and detects potential N-V qualia pairs only if the two elements are syntactically related. If the members of some qualia pairs can obviously be syntactically bound, all syntactically related pairs are not qualia pairs and, conversely, no theoretical or experimental clue ensures that qualia pairs have to be syntactically bound. Indeed, those hypotheses are partially invalidated by results of an experiment described in Section 5.4.

3. SYMBOLIC ACQUISITION OF QUALIA ELEMENTS

This section is devoted to the description of ASARES, a symbolic acquisition tool used to extract qualia pairs from corpora. It follows the previously seen 5-step approach proposed by Hearst, but its originality lies in the fact that the fourth step (detecting the common points in the examples) is considered as a machine-learning task. In order to manage this task, ASARES makes the most of a powerful symbolic machine learning technique: Inductive Logic Programming (ILP). ILP is adapted to our qualia extraction task in order to produce relevant contextual patterns (that is, from a "machine-learning" point of view, to infer rules) from examples and counter-examples of qualia pairs in the corpus.

First, the corpus used in our experiments and its several steps of tagging is presented in the next sub-section. Then, the whole machine learning process is described, including a discussion about the use of ILP for this task, the selection and the encoding of the examples, and an overview of the learning process itself.

3.1 Corpus and tagging

This sub-section is devoted to the presentation of the corpus used in our experiments. First, the choice of this corpus is described in the next sub-section. Then, Sub-sections 3.1.2 and 3.1.3 respectively present the Part-of-Speech and semantic taggings, whose information is used as a basis for the extraction patterns.

3.1.1 The MATRA-CCR Corpus

The French corpus used in this project is a 700 Kbytes handbook of helicopter maintenance, given to us by MATRA-CCR Aérospatiale, which contains more than 104,000 word occurrences. It has some specific characteristics that are especially well suited for our task: it is coherent, that is, its vocabulary and syntactic structures are homogeneous; it contains many concrete terms (screw, door, *etc.*) that are frequently used in sentences together with verbs indicating their telic ("screws must be tightened", *etc.*) or agentive roles.

3.1.2 Part-of-Speech tagging

This corpus has been tagged with Part-of-Speech (PoS) information with the help of annotation tools developed in the Multext project (Armstrong, 1996). Thus, sentences and words are first segmented with MTSEG; words are analyzed and lemmatized with MMORPH (Petitpierre and Russell, 1998; Bouillon *et al.*, 1998). Finally, words having more than one possible PoS tag are disambiguated by the TATOO tool, a Hidden Markov Model (HMM) tagger (Armstrong *et al.*, 1995), which can be trained directly on a non-disambiguated part of the corpus. Each word (eventually) receives a single tag that indicates its PoS as well as inflection information (gender, number or conjugation where it applies). Finally, the accuracy of this tagging, evaluated with a 4,000 word hand-tagged part of the corpus, is very good: only 2% of the words are detected as wrongly tagged.

3.1.3 Semantic tagging

A semantic tagging has also been performed on the corpus, following the work of Bouillon *et al.* (2000). It aims at providing some general semantic information about words (*e.g.* this word designates a human, this one an action verb, *etc.*).

The main hypotheses guiding the method of semantic tagging are that:

- morpho-syntatic information can help to distinguish meanings of words that are polyfunctional, such as *règle* in French which can be the indicative of the verb to regulate and the noun rule (see also (Wilks and Stevenson, 1996; Yarowsky, 1992; Ceusters *et al.*, 1996)),
- morpho-syntactic analysis can be done by a probabilistic (HMM) tagger and,
- more daringly, remaining semantic ambiguity can also be solved (*mutatis mutandis*) by an HMM tagger.

These hypotheses are not new, but here, we describe the way we have implemented them, and we evaluate our method with the MATRA-CCR corpus.

After the PoS-tagging and disambiguation of the corpus previously explained, one or more semantic tags are associated with each word of the corpus. The TATOO HMM tagger, applying a model trained on the ambiguous semantic tags, resolves the remaining semantic ambiguities. As we are in a restricted domain, homonyms are very rare; what need to be disambiguated here are mostly polysemes whose senses are related in a systematic way (Pustejovsky, 1995). These polysemes are particularly suitable for this kind of method since, by definition, the correct sense can be identified by the context around the word and their disambiguation does not require pragmatic disambiguation.

Thus, the first step is to choose a set of semantic tags for each category of a word. In order to classify the nouns, the most generic classes of WordNet (Fellbaum, 1998) are used. However, they are modified and refined in two ways: irrelevant classes (*i.e.* classes not used in the corpus; *e.g.* abstraction) have been withdrawn; for large classes, a more precise granularity has been chosen, in order to distinguish and characterize their elements (*e.g.* the concrete object class). This has led to 33 classes. Figure 1 presents a part of their hierarchical organization as defined in WordNet.

Concerning verbs, WordNet classification was judged too specific and divided into too many classes for our corpus. A minimal partition into five classes has been chosen: cognitive activity, physical activity, state, modality and temporality. *Ad hoc* tagsets have also been defined for all other categories of word. To sum up, here is some numerical information about the file gathering all the possible semantic tags for each word of the corpus. It contains 1489 different nouns, 129 (8.7%) of them being ambiguous (*i.e.* that can be classified in more than one class and thus receive more than one semantic tag). Most of these ambiguities correspond to complementary polysemy, in particular classical semantic alternations (for example, *enfoncement* (hollow) can both indicate a process or its result) or contextual variants (for example, *bout* (end) can be temporal or locative). The file also contains 8 different acronyms, one of them being ambiguous; 567 different verbs, 6 of them being ambiguous; 68 adjectives, 4 of them being ambiguous; 53 prepositions, 9 of them being ambiguous; about 15 determiners and 30 pronouns, none of them being ambiguous.

Each occurrence in the corpus is given all its possible tags according to this file. Then, the HMM-based disambiguation training is done just as for the PoS-tagging. However, since the ambiguities are very limited, this training has been done with a set of interesting sentences. For the evaluation,

a subset of about 6,000 words of the MATRA-CCR corpus has been manually tagged and compared with the output of the tagger. In this subset, 455 words were ambiguous (7.78%). The application of the semantic tagging method has led to a score of 1.18% of remaining errors, that is, (when compared with to the 7.78% of ambiguous words) 85% of good disambiguation.

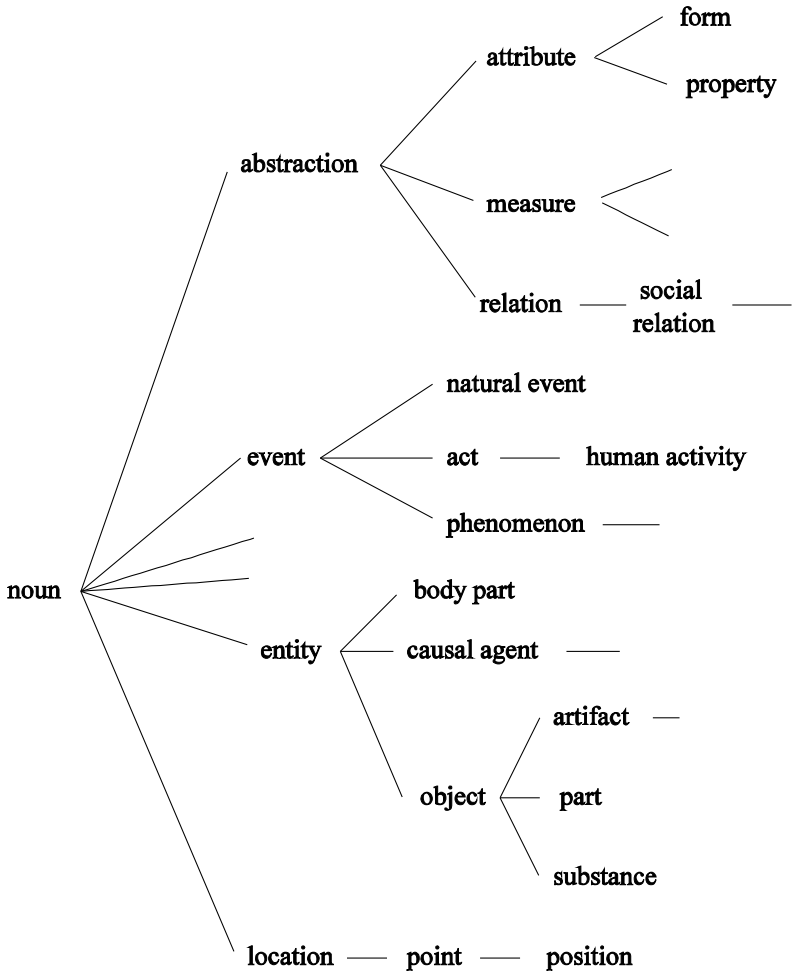

Figure #-1. Semantic hierarchy of nouns

More than one third of the remaining errors are due to prepositions. The errors concerning nouns and verbs, which are the key elements of the qualia structures we are willing to extract, are therefore relatively rare in the disambiguated corpus.

Finally, after these PoS and semantic tagging processes, a sentence such as "*L'opérateur utilise les tournevis pour visser...*" ("The operator uses the screwdrivers to screw…") appears in the following format:

m151	L'	le	det_masc_sg	-
m152	opérateur	opérateur	noun_masc_sg	human
m153	utilise	utiliser	verb_ind_3_sg	phys_action
m154	les	le	det_masc_pl	-
m155	tournevis	tournevis	noun_masc_pl	artifact
m156	pour	pour	prep	goal_prep
m157	visser	visser	verb_inf	phys_action
	…			

The first column gives a unique identifier to each word of the corpus, the second and third ones respectively contain the words as they appear and the corresponding lemmas, the fourth column gives the PoS information and the last one the semantic information.

3.2 Inferring extraction patterns with ILP

All those PoS and semantic tags in the MATRA-CCR corpus are the contextual key information used by ASARES to extract qualia pairs with the help of an inductive method called inductive logic programming (ILP). The choice of this symbolic learning method is explained in the next sub-section. Since ILP is a supervised ML technique, we need examples; the way they are obtained and their representations are described in Sub-section 3.2.2; and the learning step, which infers the extraction patterns from the examples, is finally presented in Sub-section 3.2.3.

3.2.1 About the use of ILP

Our selection of a learning method is guided by the fact that this method must not only provide a predictor (this N-V pair is qualia, this one is not), like most statistical methods, but must also infer general rules able to explain the examples, that is, give rise to linguistically interpretable elements which predict qualia relations. This essential explanatory characteristic has motivated our choice of the ILP framework (Muggleton and De Raedt, 1994) in which programs that are inferred from a set of facts (examples and counter-examples of the concept to be learned) and background knowledge, are logic programs, that is, sets of Horn clauses. Indeed, ILP's relational nature can provide a powerful expressiveness for the still unknown linguistic patterns expressing qualia relations in a given corpus. Moreover, errors inherent in the automatic PoS and semantic tagging process previously

described make the choice of an error-tolerant learning method essential. The relative ease with which ILP handles noisy data guarantees this robustness.

Most ILP systems provide a way to deal more or less with the form of the generated rules but only some of them enable a total control of this form. Moreover, the particular hierarchical structure of our PoS and semantic information makes it essential to use a relational background knowledge processing capable ILP system. For these reasons, we have chosen ALEPH (a state-of-art Prolog implementation freely available at http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/). This ILP implementation has already proved to be well suited to deal with a large amount of data in multiple domains (mutagenesis, drug structure...) and allows us to precisely customize all the settings of the learning task.

3.2.2 Example construction

As explained above, ILP algorithms generate rules explaining what characterize examples of the concept to be learned from counter-examples. In our case, we want to discriminate qualia N-V pairs from non-qualia ones according to their PoS and semantic context in the MATRA-CCR corpus. Therefore, our first task consists in building the sets of examples and counter-examples (hereafter, respectively E+ and E-), that is, in describing the sentences where qualia N-V pairs and non-qualia ones occur in terms of PoS and semantic information. It is well worth noting that no distinction is made between the different qualia roles (it is not considered as relevant for our information retrieval application of the acquired N-V pairs (Claveau and Sébillot, 2004b)); thus, every telic, agentive or formal N-V pair may be considered as an example. Here is our methodology for their construction.

Given a subset of N-V pairs of our corpus, every occurrence in the text of each pair of this subset is manually annotated as relevant or irrelevant according to Generative Lexicon's qualia structure principles. The considered occurrence is then added to the E+ set if it is annotated as relevant, to the E- one otherwise, and the contextual information of this occurrence is added to the background knowledge. The examples and counter-examples therefore contain clauses of the form is_qualia(noun identifier, verb identifier) where noun identifier and verb identifier are the unique identifier of the considered N-V pair occurrence. The contextual information is stored as background knowledge in the form of the following clauses:

```
tags(w1 identifier, PoS-tag, semantic-tag).
tags(w2 identifier, PoS-tag, semantic-tag).
```

pred(w2 identifier, w1 identifier).
tags(w3 identifier, PoS-tag, semantic-tag).
pred(w3 identifier, w2 identifier).
tags(w4 identifier, PoS-tag, semantic-tag).
pred(w4 identifier,w3 identifier).
tags(w5 identifier, PoS-tag, semantic-tag).
pred(w5 identifier, w4 identifier).
distances(w4 identifier, w2 identifier, distance in words, distance in verbs).

where, *e.g.*, the studied N-V pair w4-w2 occurs in the 5-word long sentence "*w1 w2 w3 w4 w5*", pred(x,y) indicates that word y occurs just before word x in the sentence, predicate tags/3 gives the PoS and semantic tags of a word, and distances/4 specifies the number of words and the number of verbs between N and V in the sentence. During this step, only a few word categories (determiners, some adjectives), which are not considered relevant to predicting qualia or non-qualia pairs, are not taken into account; all the other words of the sentence where the target N-V pair appears participate to its contextual description.

For example, consider the qualia pair *tournevis-visser* (screwdriver-screw) in the previously seen sentence "*L'opérateur utilise les tournevis pour visser...*" ("The operator uses the screwdrivers to screw…"). This N-V pair is indicated as being an example to ALEPH by adding the fact is_qualia(m155,m157) to the set of the examples. Contextual information about this pair is added to the background knowledge:

tags(m152,noun_masc_sg,human).
tags(m153,verb_ind_3_sg,phys_action).
pred(m153,m152).
tags(m155,noun_masc_pl,artifact).
pred(m155,m153).
tags(m156,prep,goal_prep).
pred(m156,m155).
tags(m157,verb_inf,phys_action).
pred(m157,m156).
...
distances(m155, m157, 1, 0).

About 3,000 examples and 3,000 counter-examples are automatically produced this way from the manual annotation of the qualia and non-qualia pairs in the MATRA-CCR corpus. Other information, describing the hierarchical relationships among PoS and semantic tags, is also provided in ALEPH's background. Those relationships encode, for example, the fact that a tag instrument denotes an instrument and can be considered as a kind of

artifact, which is a kind of object and so on (see Figure 1). This is easily written in the Prolog form:

```
instrument(W) :- tags(W,_,instrument).
artifact(W) :- instrument(W).
object(W) :- artifact(W).
object(W) :- part(W).
object(W) :- substance(W).
…
```

3.2.3 Learning step

In addition to the sets of examples and the various kinds of information in the background knowledge, a hypothesis language is also provided to the ILP system. It is used to precisely define the expected form of the generated rules (or hypotheses). In the qualia extraction case, this language makes the most of the PoS and semantic tags of words occurring in the examples (N-V pairs and their contexts) and distance information between N and V (a complete description of the hypothesis language used and its consequences on the learning process can be found in (Claveau *et al.*, 2003)). For example, the rules produced, which are used as patterns to extract new qualia pairs, look like:

```
is_qualia(N,V) :- precedes(N,V), near_verb(N,V), infinitive(V),
action_verb(V), artifact(N), pred(V,P), goal_preposition(P).
```

This rule means that a pair composed by a noun N and a verb V will be considered as qualia if V appears in a sentence after N, V is an action verb in the infinitive preceded by a goal preposition P and N is an artifact. Thus, this rule is equivalent to the pattern: N artifact + (any token but a verb)* + goal preposition + infinitive action verb V. This rule covers (that is, explains or logically entails) the pair *tournevis-visser* (screwdriver-screw) in the previously seen sentence "*L'opérateur doit utiliser les tournevis pour visser...*" ("The operator uses the screwdrivers to screw…") and certainly many others in the corpus.

According to the hypothesis language, the ILP algorithm infers rules that cover a maximum of examples and no counter-examples (or only a few, some noise can be allowed in order to produce more general patterns), by generalizing the examples (Muggleton and De Raedt, 1994). More precisely, the inference process follows the following steps:

1. select one example $e \in E+$ to be generalized. If none exists, stop.
2. define a hypothesis (*i.e.* potential pattern) search space H according to e and the hypothesis language;
3. search H for the rule h that maximizes a score function Sc;

4. remove the examples that are covered by the chosen rule. Return to step 1.

A precise description of the structure of the hypothesis space H, containing all the potential patterns generalizing an example, and the way it is explored to find a global optimum can be found in (Claveau *et al.*, 2003). The score function Sc depends on the number of examples and counter-examples covered by a hypothesis h, as well as its length (shorter rules are favored). Thus, the chosen rules are meaningful generalizations of the examples and reject most of the counter-examples.

This learning step, which is the heart of ASARES, takes about 15 minutes on a recent Linux PC. Several rules are produced (see next section for a detailed description) which can now be used to automatically retrieve new qualia N-V pairs in the corpus.

4. LINGUISTIC DISCUSSION ABOUT THE INFERRED PATTERNS

As mentioned previously, our choice of a symbolic ML technique is mostly motivated by the fact that ILP produces general rules or patterns that are linguistically interpretable, leading to the discovery of corpus-specific linguistic generalizations regarding the concept of qualia relation. Before analyzing the performances of the patterns inferred by ASARES in extracting qualia pairs (see Section 5), this section provides a linguistic discussion about the patterns. More precisely the question raised in Sub-section 4.1 is: what do the learned clauses tell us about the linguistic structures that are likely to convey qualia relations between a noun and a verb in the studied corpus? A comparison with manually found patterns is proposed in Sub-section 4.2.

4.1 Inferred patterns

ASARES has produced the nine following clauses from the examples and counter-examples, which we are now facing and willing to interpret linguistically:

1. is_qualia(N,V) :- precedes(V,N), near_verb(N,V), infinitive(V), action_verb(V).
2. is_qualia(N,V) :- contiguous(N,V).
3. is_qualia(N,V) :- precedes(V,N), near_word(N,V), near_verb(N,V), suc(V,X), preposition(X).
4. is_qualia(N,V) :- near_word(N,V), sentence_beginning(N).

5. is_qualia(N,V) :- precedes(N,V), singular_common_noun(N), suc(V,C), colon(C), pred(N,D), punctuation(D).
6. is_qualia(N,V) :- near_word(N,V), suc(V,C), suc(C,D), action_verb(D).
7. is_qualia(N,V) :- precedes(N,V), near_word(N,V), pred(N,C), punctuation(C).
8. is_qualia(N,V) :- near_verb(N,V), pred(V,C), pred(C,D), pred(D,E), preposition(E), sentence_beginning(N).
9. is_qualia(N,V) :- precedes(N,V), near_verb(N,V), pred(N,C), subordinating_conjunction(C).

Predicates must be read as follows: precedes(X,Y) means that X occurs somewhere in a sentence before Y. pred(X,Y) means that Y occurs immediately before X and conversely suc(Y,X) means that X occurs immediately after Y. near_word(X,Y) means that X and Y are separated by at least one word and at most 2 words, and near_verb(X,Y) that there is no verb between X and Y.

What is first striking is the fact that, at this level of generalization, few usual linguistic features remain. The clauses seem to provide very general indications and tell us very little about types of verbs (action_verb is the only information we get), nouns (common_noun) or prepositions that are likely to fit into such structures. However, the clauses contain other information, related to several aspects of linguistic descriptions, like:

- *proximity*: this is a major criterion. Most clauses indicate that the noun and the verb must be either contiguous (clause 2) or separated by at most one element (clauses 3, 4, 6, 7) and that no verb must appear between N and V (clauses 1, 3, 8, 9).
- *position*: clauses 4, 7 and 8 indicate that one of the two elements is found at the beginning of a sentence or right after a punctuation mark, whereas the relative position of N and V (precedes/2) is given in clauses 1, 3, 5, 7 and 9.
- *punctuation*: punctuation marks, and more specifically colons, are mentioned in clauses 5 and 7.
- *morpho-syntactic categorization*: the first clause detects a very important structure in the text, corresponding to action verbs in the infinitive form.

These features bring to light linguistic patterns that are very specific to the corpus, a text falling within the instructional genre. We find in this text many examples in which a verb at the infinitive form occurs at the beginning of a proposition and is followed by a noun phrase (found by clause 1). Such lists of instructions are very typical of the corpus:

- *débrancher la prise* (disconnect the plug);

- *enclencher le disjoncteur* (engage the circuit breaker);
- *déposer les obturateurs* (remove the obturators).

Clause 5, which is equivalent to the pattern V + : + (any token)* + [:,;] + singular N, highlights enumerative structures that are very frequent in the corpus, like:

- *Ouvrir : le capot coulissant, le capot droit…* (Open: the sliding cowl, the right cowl…);
- *Poser : le bouchon, la porte d'accès…* (Set: the cap, the access door…);
- *…déclenche : l'allumage du voyant 1, l'allumage du voyant alarme…* (… set up: the lighting of indicator signal 1, the lighting of alarm indicator signal…).

These results emphasize the ability of our technique to learn corpus-specific patterns. Indeed, when applied to other corpus, other experiments of qualia extraction (Claveau and Sébillot, 2004b) or close semantic relation acquisition in the Meaning-Text theory framework (Claveau and L'Homme, 2004), using the same technique have shown that most of the patterns inferred are dependent on the corpus.

4.2 Comparison to manual linguistic observations

To further evaluate these findings, we have compared the automatic learning results to linguistic observations made manually on the same corpus (Galy, 2000). É. Galy has listed a set of canonical verbal structures that convey telic information:

- infinitive V + det + N (*visser le bouchon*) (to tighten the cap)
- V + det + N (*ferment le circuit*) (close the circuit)
- N + past participle V (*bouchon maintenu*) (held cap)
- N + be + past participle V (*circuits sont raccordés*) (circuits are connected)
- N + V (*un bouchon obture*) (a cap blocks up)
- be + past participle V + par + det + N (*sont obturées par les bouchons*) (are blocked up by caps).

The two types of results show some overlap: both experiments demonstrate the significance of infinitive structures and bring to light patterns in which the verb and noun are very close to each other. Yet, the results are quite different since the learning method proposes a generalization of the structures discovered by É. Galy. In particular, the opposition between passive and active constructions is merged in clause 2 by the indication of mere contiguity (V can occur before or after N). Conversely, some clues have not been observed by manual analysis because

they are related to levels of linguistic information that are usually neglected by linguistic observation (punctuation marks and position in the sentence).

Consequently, when examining the results of the learning process from a linguistic point of view, it appears that the clauses give very general surface clues about the structures that are favored in the corpus for the expression of qualia relations. Yet, these clues are sufficient to give access to some corpus-specific patterns, which is a very interesting result.

5. EVALUATION AND COMPARISON OF PERFORMANCES

This section is devoted to various kinds of evaluation of ASARES. After a short description of the test set that makes this evaluation possible, we first present the performances of our symbolic system in qualia pair extraction. Thus, we measure the proportion of qualia pair that the nine inferred patterns detect on a test corpus manually annotated by GL experts. We then compare ASARES's results with those of various statistical extraction methods commonly used for semantic relation acquisition. We finally compare our qualia extraction system with an entirely manual approach relying on a syntactic annotation of the studied text.

5.1 Test set

To evaluate ASARES in real-world conditions, four GL experts have constructed an empirical test set. The test corpus on which the qualia-pair extraction is performed is a 32,000-word subset of the MATRA-CCR corpus. In spite of its relatively small size, it is impossible to manually examine every N-V pair to class it as qualia or non-qualia. We have thus focused our attention on seven domain relevant common nouns: *vis, écrou, porte, voyant, prise, capot, bouchon* (screw, nut, door, indicator signal, plug, cowl, cap). Of course, to prevent distortion of results, none of these common nouns were used as examples or counter-examples for the pattern induction phase by ASARES. Each N-V pair such that N is one of the seven nouns occurring within a sentence in the sub-corpus is retrieved. Then, the four experts manually tag each one as qualia or not; during this tagging phase, the eventual hyperonymic links between verbs given by our semantic tagging are not taken into account; each N-V pair is examined separately. Divergences (concerning only a few pairs) are discussed until complete agreement is reached.

Finally, among the 286 examined pairs, 66 are classified qualia (each N has between 4 and 17 V in qualia relations). This test set is therefore used to compare the extraction results of our automatic system with the human expert one.

5.2 Results of ASARES

The nine learned rules produced by ASARES have been applied to the sub-corpus. That is, each N-V pair containing one of the seven test nouns and any verb cooccurring with it within a sentence has been tested to see whether it is accepted by one of the learned rules. We present the results of this application of the patterns, and discuss the right and wrong decisions they have taken.

5.2.1 Performances

When applying inferred patterns to the corpus, we can decide to consider a N-V pair as qualia if *s* occurrences of this pair are detected in the test corpus by the learned rules, that is, if the context of the *s* occurrences correspond to the general patterns defined by the rules. Of course, if *s* is high, the precision rate is higher than if *s* is small, and conversely, for a small *s*, the recall rate is higher than for a high *s*. The recall and precision rates, measured on our test set, are thus defined (TP means True Positives, FP False Positives and FN False Negatives) according to *s*: $R(s) = TP(s) / (TP(s)+FN(s))$, $P(s) = TP(s) / (TP(s)+FP(s))$. To represent performances for every possible values of *s*, a recall-precision graph is commonly used, on which each point represents the precision of the system according to its recall for a given *s*. Figure 2 presents the recall-precision graph for ASARES on the previously described test set.

For a comparison purpose a baseline corresponding to the density of qualia couples among the N-V pairs in the sub-corpus is given; this density represents the average precision that would be obtained by a system deciding randomly whether a N-V pair is or not qualia.

In order to use ASARES, we have to choose a value for the threshold *s*. One way to do that is to choose the value that maximizes a certain quality criterion, that is, a single performance measure. We have used two measures of this kind: F-measure, the weighted harmonic mean of the R and the P, commonly defined as: $F(s) = 2P(s)R(s) / (P(s)+R(s))$, and the Φ coefficient ($\Phi(s)=((TP(s)*TN(s))-(FP(s)*FN(s))) / \sqrt{(PrP(s)*PrN(s)*AP(s)*AN(s))}$, where A = actual, Pr = predicated, P = positive, N = negative, T = true, F = false) for which a value close to 1 indicates a good result. Table 1 presents ASARES's results on our test set for the value of the threshold *s* that

maximizes the Φ coefficient (this value is equal to 1, that is, a N-V pair is considered as qualia as soon as one occurrence of this pair is covered by one of the learned rules).

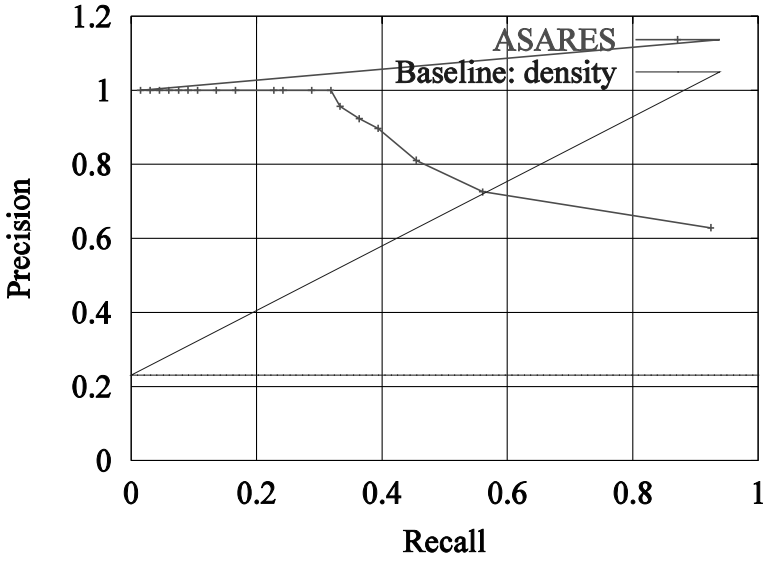

Figure #-2. Recall-precision graph

Table #-1. Optimal performances of ASARES

	Recall (%)	Precision (%)	F-measure	Φ Coefficient
ASARES	92.4	62.2	0.744	0.671

Results show a very good recall rate and a quite good precision rate. Thus, the learned rules seem to describe precisely enough the qualia concept. Such an ILP-based qualia-pair extraction system can therefore be used on the whole corpus to get relevant GL resources.

5.2.2 Extraction performances of the patterns

Before comparing ASARES's performances to those of other extraction methods, let us discuss briefly the kind of N-V pairs that are correctly retrieved, forgotten or incorrectly found using the nine patterns our system has produced. On one side, our ILP method detects most of the qualia N-V couples, like *porte-ouvrir* (door-open) or *voyant-signaler* (indicator signal-warn). The five non-detected pairs appear in very rare constructions in

our corpus, like *prise-relier* (plug-connect) in *la citerne est reliée à l'appareil par des prises* (the tank is connected to the machine by plugs) where a prepositional phrase (PP) *à l'appareil* (to the machine) is inserted between the verb and the *par*-PP (by-PP). On the other side, only 8 pairs from the 36 non-qualia pairs incorrectly detected qualia by our learning method cannot be linked syntactically. That means that the ILP algorithm can already reliably distinguish between syntactically and not syntactically linked pairs.

The main problem for ASARES is therefore to correctly identify N-V pairs related by a telic or agentive relation—the most common qualia links in our corpus—among the pairs that could be syntactically related. However, here we should carefully distinguish two types of errors. The first ones are caused by constructions that are ambiguous and where the N-V can or cannot be syntactically related, as *enlever-prises* (remove-plugs) in *enlever les shunts sur les prises* (remove the shunts from the plugs). They cannot be disambiguated by superficial clues about the context in which the V and the N occur and show the limitation of using learning only from PoS and semantic information. However, they are very rare in our corpus (8 pairs). On the contrary, all remaining errors seem more related to the parameterizing of the learning method. For example, taking into consideration the number of nouns between V and N (with the help of the hypothesis language; *cf.* Section 3.2.3) could avoid a lot of wrong pairs like *poser-capot* (put up-cover) in "*poser les obturateurs capots*" (put up cover stopcocks) or *assurer-voyant* (make sure-indicator signal) in "*s'assurer de l'allumage du voyant*" (make sure that the indicator signal is switched on).

In order to reduce the cost of manually constructing a hierarchy of semantic tags (see Section 3.1) and enhance the portability of ASARES from one corpus to another, similar learnings and evaluations have been conducted without taking into account semantic tags in the example and counter-example coding, or considering all the semantic tags except those of common nouns which is the far most populated subset. Extracted patterns and results are fully described in (Claveau, 2003), which shows that, for our corpus, semantic tagging in not that important, and especially that discarding only the most expensive noun semantic tagging leads to performances quite similar to those presented here.

5.3 Comparison with numerical methods

In order to precisely evaluate ASARES's results, we have compared the performances of the patterns it has inferred to those of common numerical extraction techniques, based on cooccurrence detection (see Section 2.1). These simple techniques are frequently used in the domain of corpus-based

collocation extraction or semantic information acquisition. In this framework, a N-V qualia pair is considered as a special kind of cooccurrence. We first present the statistical measures that we have tested, and then describe the results obtained by these techniques when applied on the same test set than ASARES. Note that our purpose is not to oppose numerical approaches to our symbolic one but rather to provide well-known baselines to interpret our results presented above.

5.3.1 Statistical measures

We have chosen twelve well-known statistical measures to carry out the qualia-pair extraction task. All of the statistical indexes we use can be expressed with the help of occurrences of N-V pairs in the corpus; a comparison of these measures, commonly used for collocation extraction tasks, can be found in (Pearce, 2002). Note that the cooccurrences of nouns and verbs are calculated in the scope of sentences and are based on the lemmas of words. With each N-V pair of the corpus, we can associate a contingency table summing up these cooccurrences as it is shown in table 2. In this table, a is the number of occurrences of the N-V pair (N_i, V_j), b of N-V pairs where the noun is N_i but the verb is not V_j, c of N-V pairs where the verb is V_j but the noun is not N_i, and d of N-V pairs where the noun is not N_i and the verb is not V_j. Let us call S the total number of N-V pair occurrences, that is, $S = a + b + c + d$.

Table #-2. Contingency table of the N-V pair (N_i - V_j)

	V_j	$V_{k,\ k \neq j}$
N_i	*a*	*b*
$N_{l,\ l \neq i}$	*c*	*d*

We can now easily express some well-known statistical association criteria such as:

- Dice coefficient (Smadja, 1993): Dice = $2a/((a+b)+(a+c))$
- Kulczinsky coefficient: Kulczinsky = $(a/2)\ ((1/(a+b)) + (1/(a+c)))$
- Ochiai coefficient: Ochiai = a / sqrt$((a+b)(a+c))$
- Mutual Information coefficient: MI = $\log_2(a / ((a+b)(a+c)))$
- Cubed Mutual Information coefficient (Daille, 1994): $MI^3 = \log_2(a^3 / ((a+b)(a+c)))$
- McConnoughy coefficient: McC = $(a^2 – bc) / ((a+b)(a+c))$
- Loglike coefficient (Dunning, 1993): Loglike = $a \log a + b \log b + c \log c + d \log d - (a + b) \log(a + b) - (a + c) \log(a + c) - (b + d) \log(b + d) - (c + d) \log(c + d) + S \log S$
- Simple matching coefficient: SMC = $(a+d) / S$

- Yule coefficient: Yule = *(ad-bc) / (ad+bc)*
- Φ^2 test (Church and Gale, 1991): Φ^2 = $(ad-bc)^2 / ((a+b)(a+c)(b+c)(b+d))$
- Cosinus (binary case): Cosinus = a / sqrt(bc)
- Jaccard coefficient (binary case): Jaccard = *a/(a+b+c)*

5.3.2 Results and discussion

All these statistical measures are then evaluated for each of the 286 N-V pairs containing one of the seven nouns. In a similar way to what we do for our ILP method, we also try to find the coefficient threshold value which maximizes the Φ coefficient for each of these statistical coefficients. Table 3 indicates the best results obtained.

Table #-3. Statistical methods results

	Recall (%)	Precision (%)	F-measure	Φ coefficient
Dice	33.3	88	0.48	0.477
Kulczinsky	36.4	70.6	0.48	0.414
Ochiai	42.4	82.4	0.56	0.517
MI	51.5	40	0.45	0.261
MI^3	36.4	92.3	0.522	0.52
McC	36.4	70.6	0.48	0.414
Loglike	42.4	80	0.554	0.505
SMC	100	25.3	0.385	0.17
Yule	53	41.2	0.464	0.279
Φ^2	37.9	78.1	0.51	0.464
Cosinus	42.4	77.8	0.549	0.493
Jaccard	31.8	87.5	0.467	0.467

One can notice that only a few statistical measures have good enough results to be used for automatic qualia pair extraction, and none of them matches the results obtained by ASARES. This is even more obvious when representing the extraction results with a recall-precision graph. Figure 3 presents such a graph with one of the statistical coefficient that achieves the best performances: MI^3 coefficient. From this graph, it appears that ASARES results are more precise (*i.e.* retrieves more real qualia pairs) than MI^3 ones, whatever the recall considered; the same result holds for the twelve statistical coefficient presented above (recall-precision graphs for the other coefficient can be found in (Claveau, 2003)).

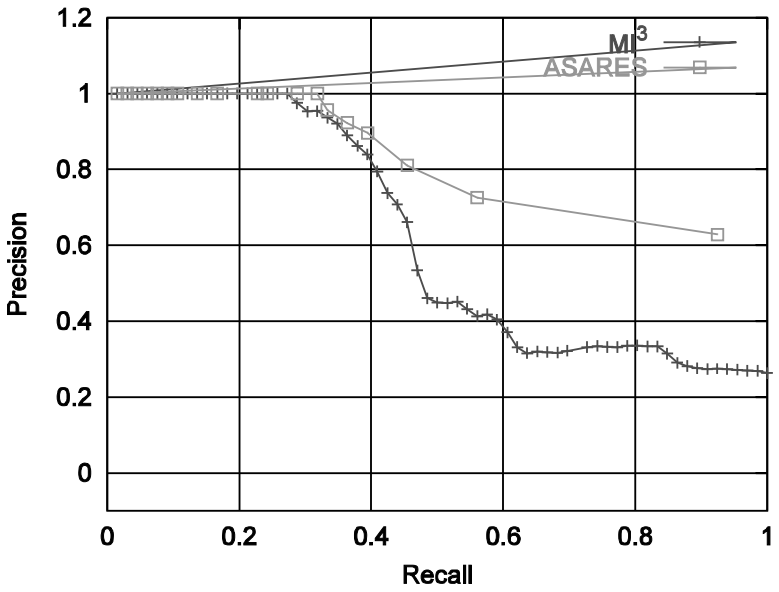

Figure #-3. Recall-precision graph of MI³ and ASARES

Of course, the differences between our ILP-based and the simple statistical-based method results can be easily explained by the differences of knowledge used by these two kinds of techniques. Indeed, while numerical models only use word lemma occurrences, our inductive learning process makes the most of PoS and semantic tags but also needs examples and counter-examples, which is a way to implicitly add linguistic knowledge to the extraction system. Nevertheless, this comparison remains interesting from a pragmatic point of view, more particularly in the balance between the choice of a supervised or unsupervised method and the resulting performances.

5.4 Comparison with a syntactic-based extraction method

We have also compared our qualia extraction system with an entirely manual approach: a syntactic annotation of the studied text. Each N-V pair occurring within a sentence of the corpus is tagged as syntactically linked (that is, the noun is subject, object, or modifier of the verb) or not. The underlying idea of this method is that a frequent syntactic link between a noun and a verb in a text may indicate a semantic link between this noun and this verb, for example a qualia link.

Therefore, a N-V pair is considered qualia if more than a certain number of its occurrences are detected syntactically linked. This threshold, as for the ILP-based and statistical methods, is chosen to maximize the Φ coefficient; the value found is 1. Table 4 gives the performances of such a system for our test set.

Table #-4. Syntactic linkedness method results

	Recall (%)	Precision (%)	F-measure	Φ coefficient
Syntactic link	86.4	79.2	0.826	0.772

These results indicate a slightly lower recall rate but a better precision rate than our ILP-based method. The fact that the recall rate is lower than 100% tends to show that, in our corpus, a qualia link is more than a basic syntactic link. Much of the 13.6% qualia pairs with non syntactically linked elements are N-V couples that appear in elliptic turns of phrases, or in which N and V are separated by a strong punctuation mark. For example, the qualia pair *voyant-allumer* (indicator signal-switch on) is not considered as syntactically related in "*éteindre le voyant ; allumer*" (switch the indicator signal off; switch on); neither is the couple *vis-poser* (screw-set) in "*poser l'ensemble : rondelle, vis et serrer au couple*" (set the whole: washer, screw and couple-tighten).

However, this better precision value seems to lead to the conclusion that our ILP-based method could improve its results, especially its precision rate, by considering syntactic information, and not only PoS and semantic ones. However, automatic syntactic annotation remains currently too noisy to be used without human supervision, and a manual annotation cannot be foreseen for a huge amount of texts. Here again, one should choose between high quality results and automatic or quasi-automatic extraction methods, accordingly to one's goals.

6. CONCLUDING REMARKS

In this chapter, we have presented ASARES, a symbolic machine learning technique that allows us to infer morpho-syntactic and semantic patterns of qualia relations from the descriptions of some pairs of Ns and Vs whose elements are linked or not by a qualia role. ASARES, more technically described in (Claveau *et al.*, 2003), produces efficient and linguistically motivated patterns, which are useful for the study of the corpus-specific structures conveying qualia roles. Those automatically obtained patterns can be applied to a corpus to successfully get GL resources and populate Generative Lexicons.

The portability of ASARES from one corpus to another may be considered as limited by two facts: the semantic tagging and the supervised nature of the method. Concerning the need for a manual semantic tagging, we have however shown (see Section 5.2.2) that semantic tags of nouns, the less portable and most expensive category, can be discarded without any loss in ASARES's performances. The second point is the manual feeding of our ILP-based system with examples and counter-examples of N-V qualia pairs. We have proved (Claveau and Sébillot, 2004a) that it is possible to combine a numeric and our symbolic approaches in a so-called semi-supervised acquisition technique in order to overcome this problem, keeping again the same performances.

Being able to acquire N-V qualia pairs with the help of ASARES has allowed us to test the relevance of N-V qualia relations in information retrieval (IR) applications, more precisely to expand users's requests to an IR system. We have automatically added qualia verbs learned on a corpus to the nouns contained in the requests, and have shown that N-V qualia expansion leads to a limited but statistically significant increase of the performances of IR systems, especially in the ranking of the first 20 documents (Claveau and Sébillot, 2004b).

Among the next steps of our research, we shall focus on N-N pairs, which very frequently exhibit telic relations in texts, as in *bouchon de protection* (protective cap). Another potential avenue is to try to learn separately each qualia semantic relation (telic, agentive, formal) instead of all together as it is done up to now. Even if such a distinction is maybe not useful for an information retrieval application, it could result in linguistically interesting rules.

References

Armstrong, S., 1996, Multext: multilingual text tools and corpora, In: *Lexikon und Text*, H. Feldweg and W. Hinrichs, ed., Tübingen:Niemeyer, pp. 107-119.

Armstrong, A., Bouillon, P., and Robert, G., 1995, *Tagger Overview*, Technical report, ISSCO, Geneva, Switzerland, http://www.issco.unige.ch/staff/robert/tatoo/ tagger.html.

Bouaud, J., Habert, B., Nazarenko, A., and Zweigenbaum, P., 1997, Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation avec deux modélisations conceptuelles, In: *Proceedings of IC'97, Ingénierie des Connaissances*, Roscoff, France, pp. 207-223.

Bouillon, P., Baud, R. H., Robert, G., and Ruch, P., 2000, Indexing by statistical tagging, In: *Proceedings of JADT'2000, Journées internationales d'analyse de données textuelles*, Lausanne, Switzerland, pp. 35-42.

Bouillon, P., Lehmann, S., Manzi, S., and Petitpierre, D., 1998, Développement de lexiques à grande échelle, In: *Proceedings of Colloque de Tunis 1997 « La mémoire des mots »*, Tunis, Tunisia, pp. 71-80.

Ceusters, W., Spyns, P., DeMoor, G., and Martin, W., 1996, *Tagging of Medical Texts: The Multi-TALE Project*, Amsterdam:IOS Press.

Church, K. W., and Gale, W. A., 1991, Concordances for parallel texts, In: *Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research*, University of Waterloo, Ontario, Canada, pp. 40-62.

Church, K. W., and Hanks, P., 1989, Word association norms, mutual information, and lexicography, In: *Proceedings of ACL'89, 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, pp. 76-83.

Claveau, V., 2003, *Acquisition automatique de lexiques sémantiques pour la recherche d'information*, PhD thesis, Université de Rennes 1, France.

Claveau, V. and L'Homme M.-C., 2004, Discovering specific relationships between nouns and verbs in a specialized French corpus. In: *Proceedings of CompuTerm'04, 3rd International Workshop on Computational Terminology*, Geneva, Switzerland.

Claveau, V., and Sébillot, P., 2004a, From efficiency to portability: acquisition of semantic relations by semi-supervised machine learning, In: *Proceedings of COLING'04, 20th International Conference on Computational Linguistics, G*eneva, Switzerland, pp. 261-267.

Claveau, V., and Sébillot, P., 2004b, Extension de requêtes par lien sémantique nom-verbe acquis sur corpus, In: *Proceedings of TALN'04, Traitement automatique des langues naturelle*, Fes, Morocco.

Claveau, V., Sébillot, P., Fabre, C., and Bouillon, P., 2003, Learning semantic lexicons from a part-of-speech and semantically tagged corpus using inductive logic programming, *Journal of Machine Learning Research, special issue on Inductive Logic Programming*, **4:**493-525.

Daille, B., 1994, *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques*, PhD thesis, Université Paris VII, France.

Dunning, T. E., 1993, Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics*, **19**(1)**:**61-74.

Fabre, C., 1996, *Interprétation automatique des séquences binominales en anglais et en français. Application à la recherche d'informations*, PhD thesis, Université de Rennes 1, France.

Fabre, C., and Sébillot, P., 1999, Semantic interpretation of binominal sequences and information retrieval, In: *Proceedings of International ICSC Congress on Computational Intelligence: Methods and Applications, CIMA'99, Symposium on Advances in Intelligent Data Analysis AIDA'99*, Rochester, N.Y., USA.

Fellbaum, C., ed., 1998, *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA.

Galy. É., 2000, *Repérer en corpus les associations sémantiques privilégiées entre le nom et le verbe : le cas de la fonction dénotée par le nom*, Master's thesis, Université de Toulouse - Le Mirail, France.

Grefenstette, G., 1997, SQLET: short query linguistic expansion techniques, palliating one-word queries by providing intermediate structure to text, In: *Proceedings of RIAO'97, Recherche d'Informations Assistée par Ordinateur*, McGill-University, Montreal, Quebec, Canada, pp. 500-509.

Grefenstette, G., 1994, *Explorations in Automatic Thesaurus Discovery*, Dordrecht: Kluwer Academic Publishers.

Harris, Z., Gottfried, M., Ryckman, T., Mattick, P. (Jr), Daladier, A., Harris, T. N., and Harris, S., 1989, *The Form of Information in Science, Analysis of Immunology Sublanguage*, Boston Studies in the Philosophy of Science, 104, Kluwer Academic Publisher, Dordrecht.

Hearst, M. A., 1992, Automatic acquisition of hyponyms from large text corpora, In: *Proceedings of COLING'92, 14th International Conference on Computational Linguistics*, Nantes, France, pp. 539-545.

Lapata, M., and Lascarides, A., 2003, A probabilisitic account of logical metonymy, *Computational Linguistics*, **29**(2)**:**263-317.

Manning, C. D., and Schütze, H., 1999, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, Massachusetts, USA.

Mitchell, T. M., 1997, *Machine Learning*, McGraw-Hill.

Muggleton, S., and De Raedt, L., 1994, Inductive logic programming: theory and methods, *Journal of Logic Programming*, **19-20:**629-679.

Oueslati, R., 1999, *Aide à l'acquisition de connaissances à partir de corpus*, PhD thesis, Université Louis Pasteur, Strasbourg, France.

Pearce, D., 2002. A comparative evaluation of collocation extraction techniques. In: *Proceedings of LREC'02, 3rd International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain.

Petitpierre, D., and Russell, G., 1998, *Mmorph - the Multext Morphology Program*, Technical report, ISSCO, Geneva, Switzerland.

Pichon, R., and Sébillot, P., 1997, *Acquisition automatique d'informations lexicales à partir de corpus : un bilan*, Research report, INRIA, N°3321, France.

Pustejovsky, J., 1995, *The Generative Lexicon*, MIT Press, Cambridge, Massachusetts, USA.

Pustejovsky, J., Bergler, S., and Anick, P., 1993, Lexical semantic techniques for corpus analysis, *Computational Linguistics*, **19**(2)**:**331-358.

Pustejovsky, J., Boguraev, B., Verhagen, M., Buitelaar, P., and Johnston, M., 1997, Semantic indexing and typed hyperlinking, In: *Proceedings of American Association for Artificial Intelligence Conference, Spring Symposium on Natural Language Processing for the World Wide Web*, Stanford, CA, USA, pp. 120-128.

Smadja F., 1993, Retrieving collocations from text: Xtract, *Computational Linguistics*, **19**(1)**:**143-178.

Vandenbroucke, L., 2000, *Indexation automatique par couples nom-verbe pertinents*, Mémoire de DES en information et documentation, Université Libre de Bruxelles, Belgium.

Wilks, Y., and Stevenson, M., 1996, *The Grammar of Sense: is Word-Sense Tagging much more than Part-of-Speech Tagging?* Technical report, University of Sheffield, UK.

Yarowsky, D., 1992, Word-sense disambiguation using statistical models of Roget's categories trained on large corpora, In: *Proceedings of COLING'92, 14th International Conference on Computational Linguistics*, Nantes, France, pp. 454-460.